Recognition of Spontaneous Conversational Speech using Long Short-Term Memory Phoneme Predictions

Martin Wöllmer, Florian Eyben, Björn Schuller, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany

[woellmer,eyben,schuller,rigoll]@tum.de

10.21437/Interspeech.2010-97

Abstract

We present a novel continuous speech recognition framework designed to unite the principles of triphone and Long Short-Term Memory (LSTM) modeling. The LSTM principle allows a recurrent neural network to store and to retrieve information over long time periods, which was shown to be well-suited for the modeling of co-articulation effects in human speech. Our system uses a bidirectional LSTM network to generate a phoneme prediction feature that is observed by a triphone-based large-vocabulary continuous speech recognition (LVCSR) decoder, together with conventional MFCC features. We evaluate both, phoneme prediction error rates of various network architectures and the word recognition performance of our Tandem approach using the COSINE database - a large corpus of conversational and noisy speech, and show that incorporating LSTM phoneme predictions in to an LVCSR system leads to significantly higher word accuracies.

Index Terms: Long Short-Term Memory, Large-Vocabulary Continuous Speech Recognition, Context Modeling, Recurrent Neural Networks

1. Introduction

The recognition performance of systems for large-vocabulary continuous speech recognition (LVCSR) heavily depends on various factors: training and evaluating systems on well-articulated, clean, read speech can lead to word error rates below 10 % [1], while disfluent and noisy speech strongly downgrades performance [2]. Thus, in recent years a large number of different strategies to cope with conversational and noisy speech has been proposed [3, 4, 5]. Most innovations can be found in the areas of speech signal preprocessing, feature enhancement, as well as speech and non-linguistic vocalization modeling (for an overview see e. g. [6]).

Apart from techniques aiming to improve the front- or back-end of automatic speech recognition (ASR) systems based on Hidden Markov Models (HMM), strategies towards improving ASR in challenging conditions by combining the HMM principle with recurrent neural networks (RNN) is an active area of research [7, 8, 9]. Generally, these techniques can be categorized into *hybrid* approaches that use RNNs for acoustic modeling while applying HMMs for decoding, and *Tandem* approaches that use the RNN output as additional features in combination with conventional (e. g. MFCC) features. However, the limitations of recurrent neural networks still prevent such hybrid or Tandem techniques from becoming a widely used standard in ASR systems. One such limitation is the so-called *vanishing gradient problem* that causes the backpropagated error in RNNs to either blow up or exponentially decay over time [10]. This strongly limits the amount of context that RNNs can access and model. Yet, due to co-articulation effects in human speech, modeling a sufficient amount of context during speech feature generation and processing is essential. On a higher level, context in speech is usually modeled via triphones and language models, while on the feature level most ASR systems incorporate only a very limited amount of context by using first and second order regression coefficients of low-level descriptors such as MFCCs as additional features.

There exist a few works that try to address the topic of considering a higher amount of context on the feature level [11] on the one hand, and solving the vanishing gradient problem in RNNs on the other hand [12, 13, 14]. An elegant and efficient way to enable long-range context modeling with recurrent neural networks has been proposed in [14] and refined in [15]: bidirectional Long Short-Term Memory (BLSTM) networks are able to model a self-learned amount of contextual information by using memory blocks in the hidden layer of RNNs. Even though this technique was shown to prevail over the triphone principle [16], phoneme modeling via BLSTM networks has so far only been investigated for the tasks of phoneme classification [15] and keyword spotting [17, 18, 19].

In this paper, we want to investigate the potential of BLSTM phoneme modeling for continuous speech recognition in a challenging conversational ASR scenario. Since previous experiments on keyword detection revealed that in the context of speech modeling via BLSTM, Tandem architectures tend to outperform hybrid approaches [20], we decided to implement a *Tandem* system that generates BLSTM phoneme predictions which are incorporated into an HMM framework. This allows us to combine Long Short-Term Memory and triphone modeling and leads to higher word accuracies when using the system for decoding continuous, noisy, and spontaneous speech as contained in the COSINE corpus [21].

The structure of this paper is as follows: Section 2 gives an overview over the COSINE corpus which we used to evaluate our system, Section 3 outlines the principle of Long Short-Term Memory (LSTM), Section 4 introduces our Tandem BLSTM-HMM architecture, and Section 5 shows experimental results.

2. The COSINE Corpus

The COnversational Speech In Noisy Environments (COSINE) corpus [21] is a relatively new database which contains multi-party conversations recorded in real world environments. The recordings were captured on a wearable recording system so that the speakers were able to walk around during recording. Since the participants were asked to speak about anything they liked and to walk to various noisy locations, the corpus consists of natural, spontaneous, and highly disfluent speaking styles partly masked by indoor and outdoor noise sources such as crowds, vehicles, and wind. The recordings were cap-

tured using multiple microphones simultaneously, however, to match most application scenarios, we exclusively used speech recorded by a close-talking microphone (Sennheiser ME-3).

We used all ten transcribed sessions, containing 11.40 hours of pairwise conversations and group discussions. All 37 speakers are fluent, but not necessarily native English speakers. Each speaker participated in only one session and the speakers' ages range from 18 to 71 years (median 21 years).

For our experiments, we used the recommended test set (sessions 3 and 10) which comprises 1.81 hours of speech. Sessions 1 and 8 were used as validation set and the remaining six session made up the training set. The vocabulary size is 4.8 k, whereas the out-of-vocabulary (OOV) rate in the test set is 3.4 %. To the best of our knowledge, there exist no benchmark ASR results for the COSINE corpus so far.

3. Long Short-Term Memory

This section briefly introduces the principle of Long Short-Term Memory networks which we use in order to generate context-sensitive phoneme predictions in our Tandem ASR system (see Section 4).

The analysis of the error flow in conventional recurrent neural nets led to the finding that long range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem [10]). This led to the introduction of Long Short-Term Memory RNNs [14]. They are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task.

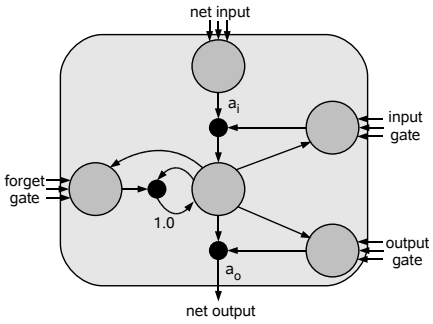

Figure 1: *LSTM memory block consisting of one memory cell: input, output, and forget gate collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state*

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative 'gate' units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Figure 1). The overall effect is to allow the network to store and retrieve information over long periods of time. For example, as long as the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

Another problem with standard RNNs is that they have access to past but not to future context. This can be overcome by using bidirectional RNNs [22], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. The amount of context information that the network actually uses is learned during training, and does not have to be specified beforehand. Figure 2 shows the structure of a simple bidirectional network.

Combining bidirectional networks with LSTM gives bidirectional LSTM [16], which has demonstrated excellent performance in many sequence labeling or pattern recognition tasks such as phoneme recognition [15], keyword spotting [17], and emotion recognition from speech [23].

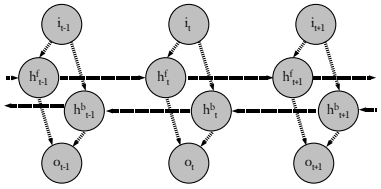

Figure 2: *Structure of a bidirectional network with input i, output o, as well as two hidden layers (h^f and h^b)*

4. Tandem BLSTM-HMM Architecture

The structure of our Tandem decoder can be seen in Figure 3: s_t and x_t represent the HMM state and the acoustic (MFCC) feature vector, respectively, while b_t corresponds to the discrete phoneme prediction of the BLSTM network (shaded nodes). Squares denote observed nodes and white circles represent hidden nodes. The HMM uses b_t as observation, in addition to the MFCC features. x_t also serves as input for the BLSTM, whereas the size of the BLSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector. The vector o_t contains one probability score for each of the P different phonemes at each time step. b_t is the index of the most likely phoneme:

$$b_t = \max_{o_t}(o_{t,1}, ..., o_{t,j}, ..., o_{t,P}) \quad (1)$$

At every time step the BLSTM generates a phoneme prediction according to Equation 1 and the HMM observes both, x_t and b_t using learned emission probabilities $p(x_t, b_t|s_t)$.

Note that the usage of bidirectional context implies a short look-ahead buffer, meaning that recognition cannot be performed truly on-line. However, for many recognition tasks it is sufficient to obtain an output e. g. at the end of an utterance, so that both, forward and backward context can be used during decoding.

5. Experiments and Results

All experiments are speaker-independent (meaning that training and testing were performed on different speakers) and were carried out using the COSINE corpus described in Section 2. As features x_t we used MFCC coefficients 1 to 12 including log. energy together with first and second order regression coefficients. To compensate for stationary noise effects, we applied cepstral mean normalization.

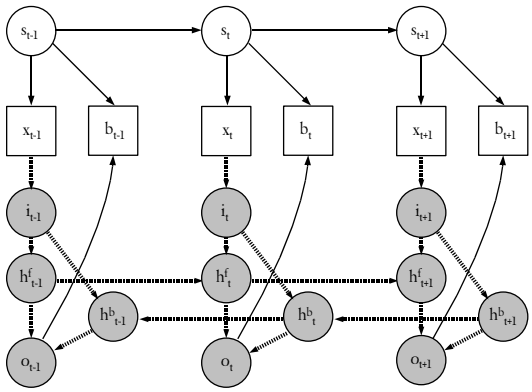

Figure 3: *Architecture of the Tandem BLSTM-HMM decoder: s_t: HMM state, x_t: acoustic feature vector, b_t: BLSTM phoneme prediction feature, i_t, o_t, h_t^f/h_t^b: input, output, and hidden nodes of the BLSTM network; squares correspond to observed nodes, white circles correspond to hidden nodes, shaded circles represent the BLSTM network.*

5.1. Network Training and Evaluation

To train and evaluate the quality of phoneme prediction, we investigated various network architectures. Since the networks were trained on framewise phoneme targets, we used an HMM system (for details see Section 5.2) to obtain phoneme borders via forced alignment. We evaluated four different network architectures: conventional recurrent neural networks, bidirectional neural networks (BRNN), unidirectional LSTM networks, and bidirectional LSTM networks. Two different variants of the respective architectures were evaluated. The first one used a single hidden layer (per input direction) composed of 128 hidden cells and memory blocks, respectively. Thereby each memory block consisted of one memory cell. The second one used three hidden layers of size 78, 128, and 80, respectively. The LSTM and BLSTM using three hidden layers per input direction consisted of one backpropagation layer (size 78) and two LSTM layers (size 128 and 80).

For training we used a learning rate of 10^{-5} and a momentum of 0.9. As a common means to improve generalization for RNNs, we added zero mean Gaussian noise with standard deviation 0.6 to the inputs during training. Prior to training, all weights were randomly initialized in the range from -0.1 to 0.1. Input and output gates used $\tanh$ activation functions, while the forget gates had logistic activation functions. Thereby we trained the networks on the standard (CMU) set of 41 different English phonemes, including targets for *silence* and *short pause*. We aborted training as soon as no improvement on the validation set (sessions 1 and 8) could be observed for at least 50 epochs, and chose the network that achieved the best framewise phoneme error rate on the validation set.

Table 1 shows the framewise error rates on the test, validation, and training set of the COSINE corpus obtained with the different network architectures. Generally, bidirectional context prevails over unidirectional context, LSTM context modeling outperforms conventional RNN architectures, and using three hidden layers leads to better performance than using only one hidden layer. The best error rate can be achieved with a BLSTM network consisting of three hidden layers (35.76 % on the validation set and 33.59 % on the test set).

network type	hidden layers	frame error rates [%] train	validation	test
BLSTM	3	23.64	35.76	**33.59**
LSTM	3	30.28	42.89	41.09
BRNN	3	48.74	50.60	49.49
RNN	3	52.37	53.11	51.09
BLSTM	1	26.79	38.16	37.02
LSTM	1	37.69	44.46	42.21
BRNN	1	51.10	51.80	50.09
RNN	1	53.17	54.64	52.85

Table 1: *Framewise phoneme error rate using the COSINE corpus and different network architectures: BLSTM, LSTM, BRNN, and RNN consisting of one and three hidden layers per input direction.*

5.2. Baseline HMM System

As explained in Section 4, we incorporate the BLSTM phoneme prediction feature into an HMM framework for LVCSR where each phoneme is represented by three emitting states (left-to-right HMMs) with 16 Gaussian mixtures. The initial monophone models consisted of one Gaussian mixture per state. All initial means and variances were set to the global means and variances of all feature vector components (flat start initialization). The monophone models were then trained using four iterations of embedded Baum-Welch re-estimation. After that, the monophones were mapped to tied-state cross-word triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for re-estimation of the triphone models. Finally, the number of mixture components of the triphone models was increased to 16 in four successive rounds of mixture doubling and re-estimation (four iterations in every round). In each round the newly created mixture components are copied from the existing ones, mixture weights are divided by two, and the means are shifted by plus and minus 0.2 times the standard deviation. Both, acoustic models and a bigram language model were trained on the training set of the COSINE corpus.

5.3. Tandem Speech Decoding

For the sake of simplicity, the BLSTM phoneme prediction feature was modeled using the same Gaussian mixture framework as for the continuous MFCC features. Since the prediction feature can be interpreted as a discrete index whose absolute value is not correlated to any intensity but rather encodes the most likely phoneme at a given time step, the weights of the Gaussians are used to represent the likelihood of a certain phoneme prediction while being in a given HMM state. By training the weights of the Gaussians, the HMM learns typical phoneme confusions of the BLSTM network that are visible as (lower weighted) Gaussian components in the respective distributions. Generally, the trained Gaussian distributions tend to form single Gaussians of low variance and high weight ('spikes') corresponding to the correct phoneme prediction in a given state as well as the most frequent confusions, and high variance Gaussians of low weight that build a 'floor value' for the phoneme predictions that are not modeled by sharp spikes in the distribution. An alternative to Gaussian mixture modeling of the phoneme predictions would be to use discrete HMMs or a mix-

network type	layers	WA [%]
BLSTM	3	45.04
LSTM	3	44.46
BRNN	3	42.59
RNN	3	43.79
BLSTM	1	44.27
LSTM	1	43.82
BRNN	1	42.95
RNN	1	43.02
baseline	-	43.36

Table 2: *Word accuracies on the COSINE test set for different Tandem models and the baseline HMM recognizer.*

ture of continuous and discrete HMMs.

Table 2 shows the word accuracies on the COSINE test set which we obtained for Tandem modeling using the different network architectures explained in Section 5.1. We can observe a similar trend as for framewise phoneme recognition (Table 1): the best performance is achieved with a Tandem model using a BLSTM network that consists of three hidden layers (word accuracy 45.04 %), leading to a significant improvement over the HMM baseline. By contrast, incorporating the phoneme predictions of a conventional RNN leads to similar, or even slightly lower word accuracies when compared to the baseline HMM.

6. Conclusion and Future Work

We proposed a system for continuous speech recognition, using phoneme predictions generated by a bidirectional Long Short-Term Memory recurrent neural network that are observed by an HMM, in addition to conventional speech features. So far, BLSTM speech modeling has only been applied for phoneme recognition, keyword spotting, and emotion recognition. In this work, we demonstrated how a combination of triphone and LSTM context modeling can be applied for noisy LVCSR. We showed that BLSTM networks can achieve a framewise phoneme recognition accuracy that significantly outperforms conventional (bidirectional) RNN architectures. When using BLSTM phoneme prediction in a Tandem manner for continuous speech recognition in a challenging spontaneous and noisy speech scenario, our Tandem model prevails over a conventional HMM system.

In the future, we will investigate *discrete* HMMs with respect to their suitability for Tandem BLSTM-HMM speech recognition. Furthermore we aim to combine the Tandem model with state-of-the-art algorithms for improving noise robustness, such as Unsupervised Spectral Subtraction, Wiener Filtering, Switching Linear Dynamic Model based feature enhancement, or Histogram Equalization.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

8. References

[1] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[2] B. Pellom and K. Hacioglu, "Recent improvements in the CU SONIC ASR system for noisy speech: the spine task," in *Proc. of ICASSP*, Hong Kong, 2003.

[3] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proc. of PIT*, Kloster Irsee, Germany, 2008, pp. 99–110.

[4] B. Mesot and D. Barber, "Switching linear dynamic systems for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1850–1858, 2007.

[5] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition - a tandem BLSTM-HMM approach," in *Proc. of Interspeech*, Brighton, UK, 2009.

[6] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement," *Journal on Audio, Speech, and Music Processing*, 2009, iD 942617.

[7] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, vol. 3, Istanbul, Turkey, 2000, pp. 1635–1638.

[8] H. Ketabdar and H. Bourlard, "Enhanced phone posteriors for improving speech recognition systems," in *IDIAP-RR*, no. 39, 2008.

[9] D. P. W. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. of ICASSP*, Salt Lake City, UT, USA, 2001, pp. 517–520.

[10] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds. IEEE Press, 2001.

[11] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. of European Conf. on Speech Communication and Technology*, Lisbon, Portugal, 2008, pp. 361–364.

[12] A. M. Schaefer, S. Udluft, and H. G. Zimmermann, "Learning long-term dependencies with recurrent neural networks," *Neurocomputing*, vol. 71, no. 13-15, pp. 2481–2488, 2008.

[13] H. Jaeger, "The echo state approach to analyzing and training recurrent neural networks," Bremen: German National Research Center for Information Technology, Tech. Rep., 2001.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[16] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Warsaw, Poland, 2005, pp. 602–610.

[17] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[18] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "A Tandem BLSTM-DBN architecture for keyword spotting with enhanced context modeling," in *Proc. of NOLISP 2009*, Vic, Spain, 2009.

[19] S. Fernandez, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proc. of ICANN*, Porto, Portugal, 2007, pp. 220–229.

[20] M. Woellmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation, Special Issue on Non-Linear and Non-Conventional Speech Processing*, 2010.

[21] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - a corpus of multi-party conversational speech in noisy environments," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, November 1997.

[23] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, 2010.